

Distributed Storage Allocations for Optimal Delay

Derek Leong

Department of Electrical Engineering
California Institute of Technology
Pasadena, California 91125, USA
derekleong@caltech.edu

Alexandros G. Dimakis

Department of Electrical Engineering
University of Southern California
Los Angeles, California 90089, USA
dimakis@usc.edu

Tracey Ho

Department of Electrical Engineering
California Institute of Technology
Pasadena, California 91125, USA
tho@caltech.edu

Abstract—We examine the problem of creating an encoded distributed storage representation of a data object for a network of mobile storage nodes so as to achieve the optimal recovery delay. A source node creates a single data object and disseminates an encoded representation of it to other nodes for storage, subject to a given total storage budget. A data collector node subsequently attempts to recover the original data object by contacting other nodes and accessing the data stored in them. By using an appropriate code, successful recovery is achieved when the total amount of data accessed is at least the size of the original data object. The goal is to find an allocation of the given budget over the nodes that optimizes the recovery delay incurred by the data collector; two objectives are considered: (i) maximization of the probability of successful recovery by a given deadline, and (ii) minimization of the expected recovery delay. We solve the problem completely for the second objective in the case of *symmetric* allocations (in which all nonempty nodes store the same amount of data), and show that the optimal symmetric allocation for the two objectives can be quite different. A simple data dissemination and storage protocol for a mobile delay-tolerant network is evaluated under various scenarios via simulations. Our results show that the choice of storage allocation can have a significant impact on the recovery delay performance, and that coding may or may not be beneficial depending on the circumstances.

I. INTRODUCTION

Consider a network of n mobile storage nodes. A source node creates a single data object of unit size (without loss of generality), and disseminates an encoded representation of it to other nodes for storage, subject to a given total storage budget T . Let x_i be the amount of coded data eventually stored in node $i \in \{1, \dots, n\}$ at the end of the data dissemination process. Any amount of data may be stored in each node, as long as the total amount of storage used over all nodes is at most the given budget T , that is, $\sum_{i=1}^n x_i \leq T$.

At some time after the completion of the data dissemination process, a data collector node begins to recover the original data object by contacting other nodes and accessing the data stored in them. We make the simplifying assumption that the stored data is instantaneously transmitted on contact; this approximates the case where there is sufficient bandwidth and time for data transmission during each contact. This data recovery process continues until the data object can be

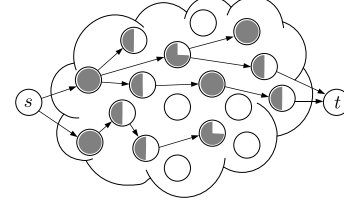


Fig. 1. Information flows originating at the source s , some of which finally arrive at the data collector t . Different amounts of coded data may eventually be stored in each storage node, subject to the given total storage budget T .

recovered from the cumulatively accessed data. Let random variable D denote the recovery delay incurred by the data collector, defined as the earliest time at which successful recovery can occur, measured from the beginning of the data recovery process. Fig. 1 depicts the information flows in such a network.

By using an appropriate code for the data dissemination process and eventual storage, successful recovery can be achieved when the total amount of data accessed by the data collector is at least the size of the original data object. This can be accomplished with random linear codes [1], [2] or a suitable MDS code, for example. Thus, if $\mathbf{r}_d \subseteq \{1, \dots, n\}$ is the set of all nodes contacted by the data collector by time d , then the recovery delay D can be written as

$$D \triangleq \min \left\{ d : \sum_{i \in \mathbf{r}_d} x_i \geq 1 \right\}.$$

Our goal is to find a storage allocation (x_1, \dots, x_n) that produces the optimal recovery delay, subject to the given budget constraint. Specifically, we shall examine the following two objectives involving the recovery delay D :

- (i) maximization of the probability of successful recovery by a given deadline d , or **recovery probability** $\mathbb{P}[D \leq d]$, and
- (ii) minimization of the **expected recovery delay** $\mathbb{E}[D]$.

By solving for the optimal allocation, we will also be able to determine whether coding is beneficial for recovery delay. For example, uncoded replication would suffice if each nonempty node is to store the data object in its entirety (i.e. $x_i \geq 1$ for all $i \in S$, and $x_i = 0$ for all $i \notin S$, where S is some subset of $\{1, \dots, n\}$); the data collector would not need to combine

This work has been supported in part by the Air Force Office of Scientific Research under grant FA9550-10-1-0166 and Caltech's Lee Center for Advanced Networking.

data accessed from different nodes in order to recover the data object.

The nodes of the network are assumed to move around and contact each other according to an exogenous random process; they are unable to change their trajectories in response to the data dissemination or recovery processes. (The recovery delay could be improved significantly if nodes were otherwise allowed to act on oracular knowledge about future contact opportunities [3], for example.)

Most work on delay-tolerant networking traditionally assume that the data object is intended for immediate consumption; both the data dissemination and recovery processes would therefore begin at the same time, and the recovery delay would be measured from the beginning of the data dissemination process. In contrast, our model more accurately reflects the characteristics of longer-term storage where the data object can be consumed long after its creation. Nonetheless, our model can still be a good approximation for short-term storage especially when the data dissemination process occurs very rapidly, as in the case of binary SPRAY-AND-WAIT [4] where the number of nodes disseminating or spraying data grows exponentially over time.

We also note that in most of the literature involving distributed storage, either the data object is assumed to be replicated in its entirety (see, for e.g., [4]), or, if coding is used, every node is assumed to store the same amount of coded data (see, for e.g., [5]–[9]). Allocations of a storage budget with nodes possibly storing different amounts of data are not usually considered.

A. Our Contribution

This paper attempts to address the gaps in our understanding of how the choice of storage allocation can affect the recovery delay performance. We formulate a simple analytical model of the problem and show that the maximization of the **recovery probability** $\mathbb{P}[D \leq d]$ can be expressed in terms of the reliability maximization problem introduced in [10]. It turns out that the simple strategies of spreading the budget minimally (i.e. uncoded replication) and spreading the budget maximally over all n nodes (i.e. assigning $x_i = \frac{T}{n}$ for all i) may both be suboptimal; in fact, the optimal allocation may not even be symmetric (we say that an allocation is *symmetric* when all nonzero x_i are equal). Applying our earlier results [11], we can show that minimal spreading is optimal among symmetric allocations when the deadline d is sufficiently small, while maximal spreading is optimal among symmetric allocations when the deadline d is sufficiently large.

For the minimization of the **expected recovery delay** $\mathbb{E}[D]$, we are able to characterize the optimal *symmetric* allocation completely: minimal spreading (i.e. uncoded replication) turns out to be optimal whenever the budget T is an integer; otherwise, the amount of spreading in the optimal symmetric allocation increases with the fractional part of T .

Interestingly, our analytical results demonstrate that the optimal symmetric allocation for the two objectives can be

quite different. In particular, when the budget T is an integer, we observe a phase transition in the optimal symmetric allocation as the deadline d increases, for the maximization of **recovery probability** $\mathbb{P}[D \leq d]$; however, minimal spreading (i.e. uncoded replication) alone turns out to be optimal for the minimization of **expected recovery delay** $\mathbb{E}[D]$.

We proceed to apply our theoretical insights to the design of a simple data dissemination and storage protocol for a mobile delay-tolerant network. Our protocol generalizes SPRAY-AND-WAIT [4] by allowing the use of variable-size coded packets. Using network simulations, we compare the performance of different symmetric allocations under various circumstances. These simulations allow us to capture the transient dynamics of the data dissemination process that were simplified in the analytical model. Our main result shows that a maximal spreading of the budget is optimal in the *high recovery probability regime*. Specifically, maximal spreading can lead to a significant reduction in the wait time required to attain a desired recovery probability. We also evaluate the protocol against a real-world data set consisting of the mobility traces of taxi cabs operating in a city. Besides validating the predictions made in our theoretical analysis, these simulations also reveal several interesting properties of the allocations under different circumstances.

B. Other Related Work

Jain et al. [12] and Wang et al. [13] evaluated the delay performance of symmetric allocations experimentally in the context of routing in a delay-tolerant network. Our results complement and generalize several aspects of their work.

We present a theoretical analysis of the problem in Section II, and undertake a simulation study in Section III. Proofs of theorems are deferred to the appendix.

II. THEORETICAL ANALYSIS

We adopt the following notation throughout the paper:

- n total number of storage nodes, $n \geq 2$
- λ contact rate between any given pair of nodes, $\lambda > 0$
- x_i amount of data stored in node $i \in \{1, \dots, n\}$, $x_i \geq 0$
- T total storage budget, $1 \leq T \leq n$
- D random variable denoting recovery delay

The indicator function is denoted by $\mathbf{I}[G]$, which equals 1 if statement G is true, and 0 otherwise. We use $\mathcal{B}(n, p)$ to denote the binomial random variable with n trials and success probability p . An allocation (x_1, \dots, x_n) is said to be *symmetric* when all nonzero x_i are equal; for brevity, let $\bar{x}(n, T, m)$ denote the symmetric allocation for n nodes that uses a total storage of T and contains exactly $m \in \{1, \dots, n\}$ nonempty nodes, that is,

$$\bar{x}(n, T, m) \triangleq \left(\underbrace{\frac{T}{m}, \dots, \frac{T}{m}}_{m \text{ terms}}, \underbrace{0, \dots, 0}_{(n-m) \text{ terms}} \right).$$

The number of contacts between any given pair of nodes in the network is assumed to follow a Poisson distribution

with rate parameter λ ; the time between contacts is therefore described by an exponential distribution with mean $\frac{1}{\lambda}$. Let W_1, \dots, W_n be i.i.d. random variables denoting the times at which the data collector first contacts node $1, \dots, n$, respectively, where $W_i \sim \text{Exponential}(\lambda)$.

A. Maximization of Recovery Probability $\mathbb{P}[D \leq d]$

Let the given recovery deadline be $d > 0$, and let the subset of nodes contacted by the data collector by time d be $\mathbf{r} \subseteq \{1, \dots, n\}$. Successful recovery occurs by time d if and only if the total amount of data stored in the subset \mathbf{r} of nodes is at least 1. In other words, the recovery delay D is at most d if and only if $\sum_{i \in \mathbf{r}} x_i \geq 1$. Since the data collector contacts each node by time d independently with constant probability $p_{\lambda,d}$, given by

$$p_{\lambda,d} \triangleq \mathbb{P}[W \leq d] = F_W(d) = 1 - e^{-\lambda d},$$

it follows that the probability of contacting exactly a subset \mathbf{r} of nodes by time d is $p_{\lambda,d}^{|\mathbf{r}|} (1 - p_{\lambda,d})^{n-|\mathbf{r}|}$. The recovery probability $\mathbb{P}[D \leq d]$ can therefore be obtained by summing over all possible subsets \mathbf{r} that allow successful recovery:

$$\mathbb{P}[D \leq d] = \sum_{\substack{\mathbf{r} \subseteq \{1, \dots, n\}: \\ |\mathbf{r}| \geq 1}} p_{\lambda,d}^{|\mathbf{r}|} (1 - p_{\lambda,d})^{n-|\mathbf{r}|} \cdot \mathbf{I} \left[\sum_{i \in \mathbf{r}} x_i \geq 1 \right]. \quad (1)$$

We seek an optimal allocation (x_1, \dots, x_n) of the budget T (that is, subject to $\sum_{i=1}^n x_i \leq T$, where $x_i \geq 0$ for all i) that maximizes $\mathbb{P}[D \leq d]$, for a given choice of n, λ, d , and T .

This problem matches the reliability maximization problem of [11] with $p_{\lambda,d}$ as the access probability; we recall that the optimal allocation may be nonsymmetric and can be difficult to find. However, if we restrict the optimization to only *symmetric* allocations, then we can specify the solution for a wide range of parameter values of $p_{\lambda,d}$ and T . Specifically, if λ or d is sufficiently small, e.g. $p_{\lambda,d} \leq \frac{1}{\lceil \frac{n}{T} \rceil}$, then $\bar{\mathbf{x}}(n, T, m = \lceil \frac{n}{T} \rceil)$, which corresponds to a minimal spreading of the budget (i.e. uncoded replication), is an optimal symmetric allocation. On the other hand, if λ or d is sufficiently large, e.g. $p_{\lambda,d} \geq \frac{4}{3\lceil \frac{n}{T} \rceil}$, then either $\bar{\mathbf{x}}(n, T, m = \lfloor \frac{n}{T} \rfloor)$ or $\bar{\mathbf{x}}(n, T, m = n)$, which correspond to a maximal spreading of the budget, is an optimal symmetric allocation.

B. Minimization of Expected Recovery Delay $\mathbb{E}[D]$

Rewriting (1) in terms of the underlying random variables gives us the following c.d.f. for the recovery delay D :

$$F_D(t) = \sum_{\substack{\mathbf{r} \subseteq \{1, \dots, n\}: \\ |\mathbf{r}| \geq 1}} (F_W(t))^{|\mathbf{r}|} (1 - F_W(t))^{n-|\mathbf{r}|} \cdot \mathbf{I} \left[\sum_{i \in \mathbf{r}} x_i \geq 1 \right].$$

Differentiating $F_D(t)$ wrt t produces the p.d.f.

$$f_D(t) = \sum_{\substack{\mathbf{r} \subseteq \{1, \dots, n\}: \\ |\mathbf{r}| \geq 1}} (F_W(t))^{|\mathbf{r}|-1} (1 - F_W(t))^{n-|\mathbf{r}|} (|\mathbf{r}| - n F_W(t)) f_W(t) \cdot \mathbf{I} \left[\sum_{i \in \mathbf{r}} x_i \geq 1 \right].$$

Therefore, assuming $\sum_{i=1}^n x_i \geq 1$ which is necessary for successful recovery, we can compute the expected recovery delay as follows:

$$\begin{aligned} \mathbb{E}[D] &= \int_0^\infty t f_D(t) dt \\ &= \sum_{\substack{\mathbf{r} \subseteq \{1, \dots, n\}: \\ |\mathbf{r}| \geq 1}} \left(\int_0^\infty t (F_W(t))^{|\mathbf{r}|-1} (1 - F_W(t))^{n-|\mathbf{r}|} (|\mathbf{r}| - n F_W(t)) f_W(t) dt \right) \cdot \mathbf{I} \left[\sum_{i \in \mathbf{r}} x_i \geq 1 \right] \\ &= \frac{1}{\lambda} \left(H_n - \sum_{\substack{\mathbf{r} \subseteq \{1, \dots, n\}: \\ 1 \leq |\mathbf{r}| \leq n-1}} \frac{1}{(n-|\mathbf{r}|) \binom{n}{|\mathbf{r}|}} \cdot \mathbf{I} \left[\sum_{i \in \mathbf{r}} x_i \geq 1 \right] \right), \end{aligned} \quad (2)$$

where $H_n \triangleq \sum_{i=1}^n \frac{1}{i}$ is the n^{th} harmonic number. We seek an optimal allocation (x_1, \dots, x_n) of the budget T (that is, subject to $\sum_{i=1}^n x_i \leq T$, where $x_i \geq 0$ for all i) that minimizes $\mathbb{E}[D]$, for a given choice of n, λ , and T . Note that the optimal allocation is independent of λ for the minimization of $\mathbb{E}[D]$ but not for the maximization of $\mathbb{P}[D \leq d]$.

The optimal value of $\mathbb{E}[D]$ can be bounded as follows:

Lemma 1. *The expected recovery delay $\mathbb{E}[D]$ of an optimal allocation is at least*

$$\frac{1}{\lambda} \left(H_n - \sum_{r=1}^{n-1} \frac{\min(\frac{rT}{n}, 1)}{n-r} \right).$$

We make the following conjecture about the optimal allocation, based on our numerical observations:

Conjecture. *A symmetric optimal allocation always exists for any n, λ , and T .*

As a simplification, we now proceed to restrict the optimization to only *symmetric* allocations (which are easier to describe and implement, and appear to perform well). For the symmetric allocation $\bar{\mathbf{x}}(n, T, m)$, successful recovery occurs by a given deadline d if and only if $\lceil 1 / (\frac{T}{m}) \rceil = \lceil \frac{m}{T} \rceil$ or more nonempty nodes are contacted by the data collector by time d , out of a total of m nonempty nodes. It follows that the resulting recovery probability is given by $\mathbb{P}[D \leq d] = \mathbb{P}[\mathcal{B}(m, p_{\lambda,d}) \geq \lceil \frac{m}{T} \rceil]$. We therefore obtain the following c.d.f. and p.d.f. for the recovery delay D :

$$\begin{aligned} F_D(t) &= \sum_{r=\lceil \frac{m}{T} \rceil}^m \binom{m}{r} (F_W(t))^r (1 - F_W(t))^{m-r}, \\ f_D(t) &= \binom{m}{\lceil \frac{m}{T} \rceil} \left(\frac{m}{T} \right) (F_W(t))^{\lceil \frac{m}{T} \rceil - 1} (1 - F_W(t))^{m - \lceil \frac{m}{T} \rceil} f_W(t). \end{aligned}$$

Thus, we can compute the expected recovery delay as follows:

$$\mathbb{E}[D] = \int_0^\infty t f_D(t) dt = \frac{1}{\lambda} \sum_{i=1}^{\lceil \frac{m}{T} \rceil} \frac{1}{m - \lceil \frac{m}{T} \rceil + i} \triangleq E_D(\lambda, T, m).$$

Fig. 2 compares the performance of different symmetric allocations over different budgets T , for an instance of n and λ ; the value of m corresponding to the optimal symmetric allocation appears to change in a nontrivial manner as we vary the budget T . Fortunately, we can eliminate many candidates for

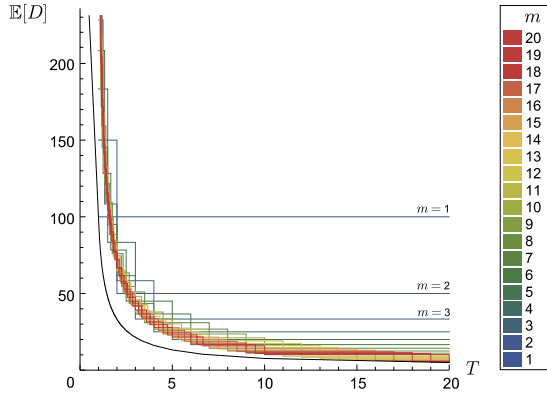


Fig. 2. Plot of expected recovery delay $\mathbb{E}[D]$ against budget T for each symmetric allocation $\bar{x}(n, T, m)$, for $(n, \lambda) = (20, \frac{1}{100})$. Parameter m denotes the number of nonempty nodes in the symmetric allocation. The black curve gives a lower bound for the expected recovery delay of an optimal allocation, as derived in Lemma 1.

the optimal value of m by making the following observation (a similar observation was made in the maximization of the recovery probability [11]): For fixed n , λ , and T , we have

$$\left\lceil \frac{m}{T} \right\rceil = k \quad \text{when } m \in ((k-1)T, kT],$$

for $k = 1, 2, \dots, \lfloor \frac{n}{T} \rfloor$, and finally,

$$\left\lceil \frac{m}{T} \right\rceil = \left\lfloor \frac{n}{T} \right\rfloor + 1 \quad \text{when } m \in \left(\left\lfloor \frac{n}{T} \right\rfloor T, n \right].$$

Since $\frac{1}{\lambda} \sum_{i=1}^k \frac{1}{m-k+i}$ is decreasing in m for constant λ and k , it follows that $E_D(\lambda, T, m)$ is minimized over each of these intervals of m when we pick m to be the largest integer in the corresponding interval. Thus, given n , λ , and T , we can find an optimal m^* that minimizes $E_D(\lambda, T, m)$ over all m from among $\lfloor \frac{n}{T} \rfloor$ candidates:

$$\{ \lfloor T \rfloor, \lfloor 2T \rfloor, \dots, \left\lfloor \left\lfloor \frac{n}{T} \right\rfloor T \right\rfloor, n \}. \quad (3)$$

Note that when $m = \lfloor kT \rfloor$, $k \in \mathbb{Z}^+$, the expected recovery delay simplifies to the following expression:

$$E_D(\lambda, T, m = \lfloor kT \rfloor) = \frac{1}{\lambda} \sum_{i=1}^k \frac{1}{\lfloor kT \rfloor - k + i}.$$

By further eliminating suboptimal candidate values for m^* using suitable bounds for the harmonic number, we are able to completely characterize the optimal symmetric allocation for any n , λ , and T :

Theorem 1. Suppose $T = a + 1 - \frac{1}{\ell}$, where $a \in \mathbb{Z}^+$, $\ell \geq 1$. If $\lfloor \ell \rfloor \leq \lfloor \frac{n}{T} \rfloor$, then

$$\bar{x}(n, T, m = \lfloor \lfloor \ell \rfloor T \rfloor)$$

is an optimal symmetric allocation; if $\lfloor \ell \rfloor > \lfloor \frac{n}{T} \rfloor$, then

$$\text{either } \bar{x}(n, T, m = \lfloor \lfloor \frac{n}{T} \rfloor T \rfloor) \text{ or } \bar{x}(n, T, m = n)$$

is an optimal symmetric allocation.

If the budget T is an integer (i.e. $\ell = 1$), then $\lfloor \ell \rfloor \leq \lfloor \frac{n}{T} \rfloor$ is always true, and so $\bar{x}(n, T, m = \lfloor T \rfloor)$, which corresponds to a

minimal spreading of the budget (i.e. uncoded replication), is an optimal symmetric allocation. However, if the budget T is not an integer (i.e. $\ell > 1$), then the amount of spreading in the optimal symmetric allocation increases with the fractional part of T , up to a point at which either $\bar{x}(n, T, m = \lfloor \lfloor \frac{n}{T} \rfloor T \rfloor)$ or $\bar{x}(n, T, m = n)$, which correspond to a maximal spreading of the budget, becomes optimal. Minimal spreading (i.e. uncoded replication) therefore performs well over the whole range of budgets T , being optimal among symmetric allocations whenever T is an integer (its suboptimality at noninteger $T = T_0$ can be bounded by the step difference in $E_D(\lambda, T, m = \lfloor T \rfloor)$ between $T = T_0$ and $T = \lceil T_0 \rceil$, since $E_D(\lambda, T, m)$ is a non-increasing function of T).

In summary, we note that the optimal symmetric allocation for the two objectives can be quite different. In particular, when the budget T is an integer, we observe a phase transition from a regime where minimal spreading is optimal to a regime where maximal spreading is optimal, as the deadline d increases, for the maximization of **recovery probability** $\mathbb{P}[D \leq d]$; however, with the averaging over both regimes, minimal spreading (i.e. uncoded replication) alone turns out to be optimal for the minimization of **expected recovery delay** $\mathbb{E}[D]$.

III. SIMULATION STUDY

We apply our theoretical insights to the design of a simple data dissemination and storage protocol for a mobile delay-tolerant network. Our protocol extends SPRAY-AND-WAIT [4] by allowing nodes to store *coded* packets that are each $\frac{1}{w}$ the size of the original data object, where parameter w is a positive integer; successful recovery occurs when the data collector accesses at least w such packets. Different *symmetric* allocations of the given total storage budget T can be realized by choosing different values of w ; the original protocol, which uses uncoded replication, corresponds to $w = 1$.

A. Protocol Description

The source node begins with a total storage budget of T times the size of the original data object, which translates to wT coded packets, each $\frac{1}{w}$ the size of the original data object. Whenever a node with more than one packet contacts another node without any packets, the former gives *half its packets* to the latter. The actual amount of data stored or transmitted by a node never exceeds the size of the original data object (or w packets) since the excess packets can always be generated on demand (using random linear coding, for example). To reduce the total transmission cost incurred, a node can also directly transmit *one packet* to each node it meets when it has w or fewer packets left; otherwise, these last few packets would be transmitted multiple times by different nodes. The dissemination process is completed when no node has more than one packet.

B. Network Model and Simulation Setup

We implemented a discrete-time simulation of $n = 100$ wireless mobile nodes in a 1000×1000 grid. A random way-point mobility model is assumed where at each time step, each

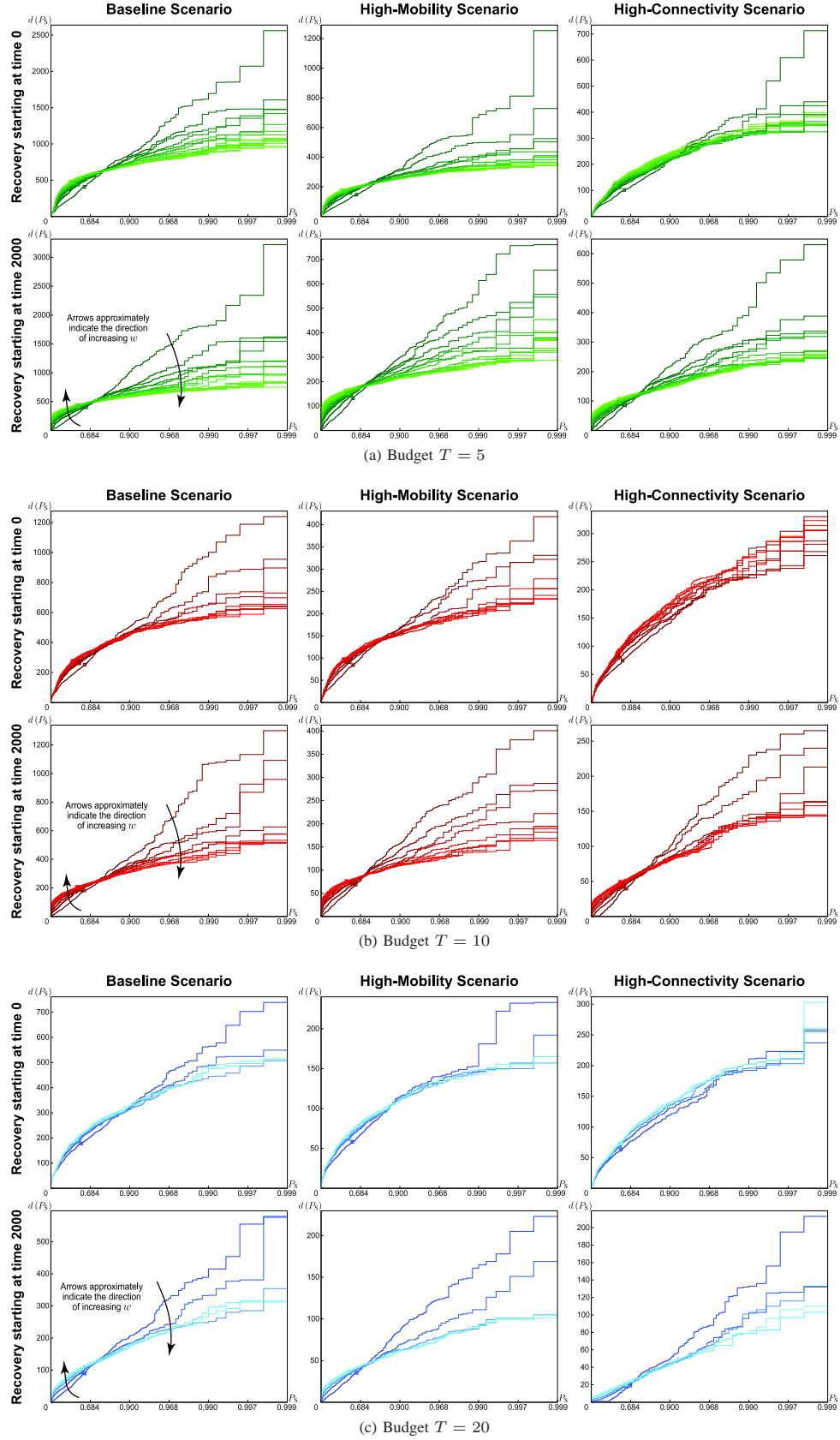


Fig. 3. (Random Waypoint) Plots of required wait time $d(P_S)$ against desired recovery probability P_S (semilogarithmic-scale), for budgets $T = 5, 10, 20$. Each colored line represents a specific choice of parameter $w \in \{1, \dots, \frac{n}{T}\}$, with $w = 1$ (darkest) corresponding to a minimal spreading of the budget (i.e. uncoded replication), and $w = \frac{n}{T}$ (lightest) corresponding to a maximal spreading of the budget. The mean recovery delay corresponding to each line is indicated by a square marker.

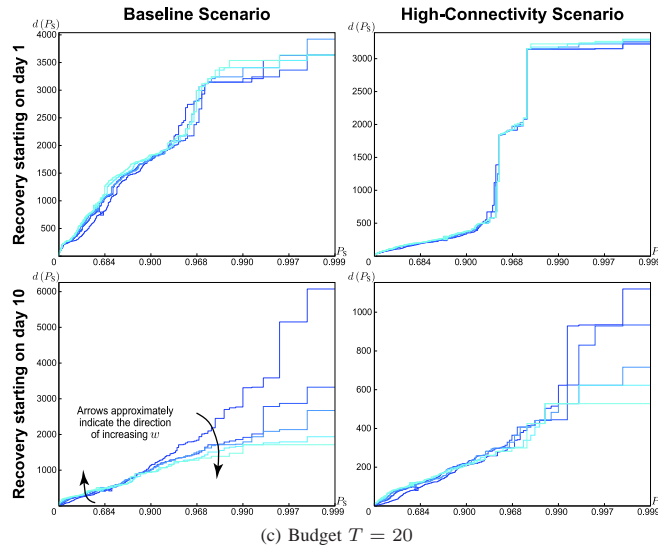
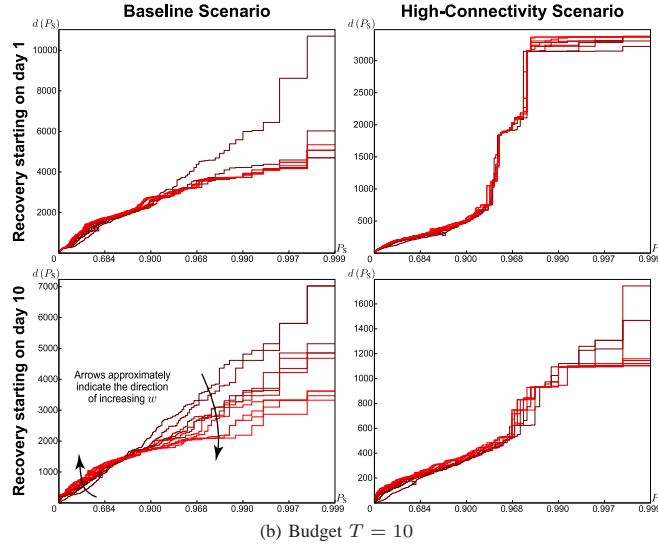
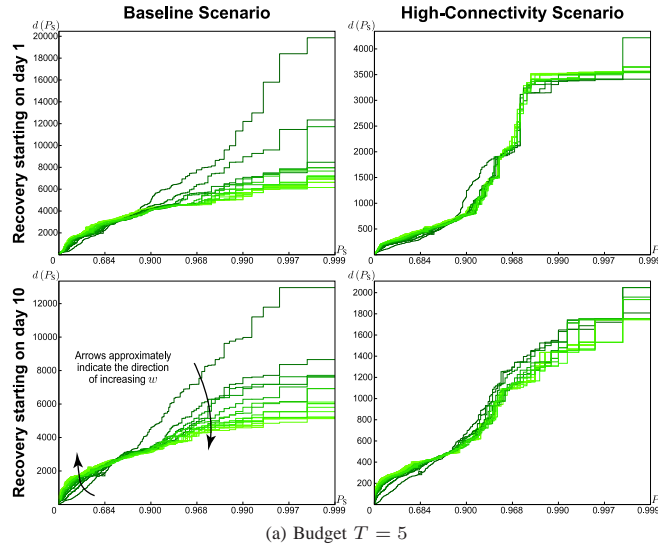


Fig. 4. (Mobility Traces) Plots of required wait time in minutes $d(P_S)$ against desired recovery probability P_S (semilogarithmic-scale), for budgets $T = 5, 10, 20$. Each colored line represents a specific choice of parameter $w \in \{1, \dots, \frac{n}{T}\}$, with $w = 1$ (darkest) corresponding to a minimal spreading of the budget (i.e. uncoded replication), and $w = \frac{n}{T}$ (lightest) corresponding to a maximal spreading of the budget. The mean recovery delay corresponding to each line is indicated by a square marker.

node moves a random distance $L \sim \text{Uniform}[5,10]$ towards a selected destination; on arrival, the node selects a random point on the grid as its next destination. Each node has a communication range of 20, and the bandwidth of each point-to-point link is large enough to support the transmission of w packets in one time step. At each time step, a maximal number of transmissions are randomly scheduled such that each node can transmit to or receive from at most one other node in range, and exactly one node may transmit in the range of a node receiving a transmission. In addition to this **baseline scenario**, we also considered the following two scenarios:

- (i) a **high-mobility scenario**, where the distance traveled by each node is increased to $L \sim \text{Uniform}[25,50]$, and
- (ii) a **high-connectivity scenario**, where the communication range is increased to 80.

We measured the recovery delay incurred by the data collector for two cases:

- (i) when the data recovery process begins at **time 0**, i.e. at the beginning of the data dissemination process, and
- (ii) when the data recovery process begins at **time 2000**, i.e. when the data dissemination process is already underway or completed. (This is a more appropriate performance metric for longer-term storage.)

We ran the simulation 500 times for each choice of budget $T \in \{5, 10, 20\}$ and parameter $w \in \{1, 2, \dots, \frac{n}{T}\}$ under each scenario, with a random pair of nodes appointed as the source and data collector for each run.

C. Simulation Results

Fig. 3 shows how the required wait time $d(P_S)$, given by

$$d(P_S) \triangleq \min\{d : \mathbb{P}[D \leq d] \geq P_S\},$$

varies with the desired recovery probability P_S for each choice of parameter w ; these plots essentially describe how much time must elapse before a desired percentage of data collectors are able to recover the data object. The **recovery probability** performance of the protocol (which can be inferred by flipping the axes) is mostly consistent with our analysis in Section II-A; specifically, the phase transition in the optimal symmetric allocation is clearly discernible in most of the plots. The **expected recovery delay** performance is also mostly consistent with our analysis in Section II-B, with minimal spreading of the budget ($w = 1$) being optimal in most of the plots.

The plots for the high-mobility scenario appear to be vertically scaled versions of the plots for the baseline scenario. This is not surprising because an increase in node mobility approximately translates to a speeding up of time. The effect of increasing node connectivity, on the other hand, seems less straightforward: the phase transition in the optimal symmetric allocation is evident for recovery starting at time 2000 but not for recovery starting at time 0. This discrepancy suggests that the data dissemination process is somewhat impeded by the increased connectivity, possibly due to greater interference.

We observe that in the *high recovery probability regime*, maximal spreading of the budget ($w = \frac{n}{T}$) can lead to a

significant reduction in the required wait time. For example, given a budget of $T = 10$ and a desired recovery probability of $P_S = 0.99$, choosing maximal spreading ($w = 10$) instead of minimal spreading or uncoded replication ($w = 1$) can yield a reduction of 40% to 60% in the required wait time for the baseline and high-mobility scenarios.

We also observe that the recovery start time appears to have a limited impact on how the different allocations perform relative to each other; the most noticeable effect of starting recovery at time 0 is the reduced spread in performance across different choices of parameter w , especially in the low recovery probability regime. This can be explained by the similarity of the different allocations during the data dissemination process: in the beginning, the different choices of parameter w would see the same allocation of the budget over the nodes because only a few nodes have been reached by the source directly or indirectly through relays; the different allocations are eventually realized only after a sufficient amount of time has passed.

D. Evaluation on Mobility Traces

To gain a better understanding of how our protocol might perform in a real-world setting, we evaluated it on a CRAWDAD data set comprising mobility traces of taxi cabs in San Francisco [14]. The traces of 100 randomly selected cabs with GPS coordinate readings over the span of an 18-day period were used. The GPS readings were sampled at approximately 60-second intervals; because reading times were not synchronized across cabs, we estimated the position of a cab at any given time using linear interpolation. For better accuracy, we assumed that a cab became inactive whenever the time between consecutive readings exceeded 2 minutes. As in the preceding simulations, we considered different scenarios and data recovery start times. Two scenarios were considered here:

- (i) a **baseline scenario**, where the communication range of each cab is 20 m, and
- (ii) a **high-connectivity scenario**, where the communication range is increased to 80 m.

We measured the recovery delay incurred by the data collector for two cases:

- (i) when the data recovery process begins on **day 1**, and
- (ii) when the data recovery process begins on **day 10**, i.e. half-way through the 18-day period.

We ran the simulation 500 times for each choice of budget $T \in \{5, 10, 20\}$ and parameter $w \in \{1, 2, \dots, \frac{n}{T}\}$ under each scenario, with a random pair of cabs appointed as the source and data collector for each run.

Fig. 4 shows how the required wait time $d(P_S)$ varies with the desired recovery probability P_S for each choice of parameter w . Compared to the plots of Fig. 3 for the random waypoint simulations, these plots exhibit distinct “jumps” in the wait times, which can be attributed to the reduced mobility of the cabs at night. Despite these nonideal conditions, many of the observations made for the previous simulations are

still applicable here. For instance, the phase transition in the optimal symmetric allocation is discernible in most of the plots for the baseline scenario. Also, starting recovery on day 1 has the effect of reducing the spread in performance across different choices of parameter w , especially in the low recovery probability regime.

Once again, we observe that in the *high recovery probability regime*, maximal spreading of the budget ($w = \frac{n}{T}$) can lead to a significant reduction in the required wait time. For example, given a budget of $T = 10$ and a desired recovery probability of $P_S = 0.99$, choosing maximal spreading ($w = 10$) instead of minimal spreading or uncoded replication ($w = 1$) can yield a reduction of 30% to 50% in the required wait time for the baseline scenario.

IV. CONCLUSION

We examined the recovery delay performance of different distributed storage allocations for a network of mobile storage nodes. Our theoretical analysis and simulation study show that the choice of objective function (i.e. recovery probability vs expected recovery delay) can lead to very different optimal symmetric allocations, and that picking the right allocation for the given circumstances can make a significant difference in performance.

The work in this paper can be extended in several directions. The simple contact model assumed here can be generalized to the case where a variable amount of data is transmitted during each contact between nodes. Another natural generalization is to allow nonuniform contact rates λ_i between the data collector and individual nodes.

APPENDIX PROOFS OF THEOREMS

Proof of Lemma 1: Consider a feasible allocation (x_1, \dots, x_n) ; we have $\sum_{i=1}^n x_i \leq T$, where $x_i \geq 0$, $i = 1, \dots, n$. Let S_r denote the number of r -subsets of $\{x_1, \dots, x_n\}$ that have a sum of at least 1, where $r \in \{1, \dots, n\}$. Recall from Lemma 1 in [11] that S_r can be bounded as follows:

$$S_r \leq \min \left(\binom{n-1}{r-1} T, \binom{n}{r} \right).$$

We can now rewrite (2) in terms of S_r by enumerating subsets according to size:

$$\begin{aligned} \mathbb{E}[D] &= \frac{1}{\lambda} \left(H_n - \sum_{r=1}^{n-1} S_r \cdot \frac{1}{(n-r) \binom{n}{r}} \right) \\ &\geq \frac{1}{\lambda} \left(H_n - \sum_{r=1}^{n-1} \frac{\min \left(\binom{n-1}{r-1} T, \binom{n}{r} \right)}{(n-r) \binom{n}{r}} \right) \\ &= \frac{1}{\lambda} \left(H_n - \sum_{r=1}^{n-1} \frac{\min \left(\frac{rT}{n}, 1 \right)}{n-r} \right). \end{aligned}$$

Proof of Theorem 1: Suppose $T = a + 1 - \frac{1}{\ell}$, where $a \in \mathbb{Z}^+$, $\ell \geq 1$. Since $kT = (a+1)k - \frac{k}{\ell}$, the expected recovery delay for the symmetric allocation $\bar{x}(n, T, m = \lfloor kT \rfloor)$, where $k \in \mathbb{Z}^+$, can be written as

$$\begin{aligned} E_D(\lambda, T, m = \lfloor kT \rfloor) &= \frac{1}{\lambda} \sum_{i=1}^k \frac{1}{(a+1)k - \lceil \frac{k}{\ell} \rceil - k + i} \\ &= \frac{1}{\lambda} \sum_{i=1}^k \frac{1}{ak - \lceil \frac{k}{\ell} \rceil + i}. \end{aligned}$$

Observe that $\lceil \frac{k}{\ell} \rceil = v$ when $k \in ((v-1)\ell, v\ell]$, for $v = 1, 2, \dots$. To compare $E_D(\lambda, T, m = \lfloor kT \rfloor)$ within each of these intervals of k , we introduce Lemma 2:

Lemma 2. For $a, v, k \in \mathbb{Z}^+$, $k \geq \frac{v}{a}$, the function

$$f(a, v, k) \triangleq \sum_{i=1}^k \frac{1}{ak - v + i} = H_{ak-v+k} - H_{ak-v}$$

decreases with k .

Proof of Lemma 2: Let $\Delta(a, v, k)$ denote the difference in the function value between consecutive values of k , that is,

$$\begin{aligned} \Delta(a, v, k) &\triangleq f(a, v, k) - f(a, v, k+1) \\ &= (H_{ak-v+k} - H_{ak-v}) - (H_{ak-v+k+a+1} - H_{ak-v+a}) \\ &= (H_{ak-v+a} - H_{ak-v}) - (H_{ak-v+k+a+1} - H_{ak-v+k}) \\ &= \left(\sum_{i=1}^a \frac{1}{ak-v+i} - \frac{1}{ak-v+k+i} \right) - \frac{1}{ak-v+k+a+1} \\ &= \left(\sum_{i=1}^a \frac{k}{(ak-v+i)(ak-v+k+i)} \right) - \frac{1}{ak-v+k+a+1}. \end{aligned}$$

We will proceed to show that $\Delta(a, v, k) > 0$ for any $a, v, k \in \mathbb{Z}^+$, $k \geq \frac{v}{a}$. First, we find a lower bound for the summation term using a geometrical argument. Consider the function

$$g(t) \triangleq \frac{k}{(ak-v+t)(ak-v+k+t)},$$

which has the second derivative

$$g''(t) = \frac{2}{(ak-v+t)^3} - \frac{2}{(ak-v+k+t)^3}.$$

For any $a, v, k \in \mathbb{Z}^+$, $k \geq \frac{v}{a}$, the function $g(t)$ is positive, decreasing with t , and convex (since $g''(t) > 0$), on the interval $t \in (0, \infty)$. We therefore have the lower bound

$$\sum_{i=1}^a \frac{k}{(ak-v+i)(ak-v+k+i)} > \int_1^{a+1} g(t) dt + \frac{g(1) - g(a+1)}{2},$$

which implies that

$$\begin{aligned} \Delta(a, v, k) &> \ln \left(\frac{(ak-v+a+1)(ak-v+k+1)}{(ak-v+k+a+1)(ak-v+1)} \right) \\ &\quad + \frac{k}{2(ak-v+1)(ak-v+k+1)} \\ &\quad - \frac{k}{2(ak-v+a+1)(ak-v+k+a+1)} \\ &\quad - \frac{1}{ak-v+k+a+1} \triangleq h(a, v, k). \end{aligned}$$

Now, it suffices to show that $h(a, v, k) \geq 0$ for any $a, v, k \in \mathbb{Z}^+$, $k \geq \frac{v}{a}$. This is indeed the case since

$$\lim_{k \rightarrow \infty} h(a, v, k) = 0,$$

and the partial derivative $\frac{\partial}{\partial k} h(a, v, k)$, which is given by

$$\frac{a}{2} \left(\frac{2(ak - v + a + 1) + 1}{(ak - v + a + 1)^2} - \frac{2(ak - v + 1) + 1}{(ak - v + 1)^2} \right) + \frac{a + 1}{2} \left(\frac{2(ak - v + k + 1) + 1}{(ak - v + k + 1)^2} - \frac{2(ak - v + k + a + 1) - 1}{(ak - v + k + a + 1)^2} \right),$$

can be shown to be negative. ■

It follows from Lemma 2 that for each $v \in \mathbb{Z}^+$, the expected recovery delay $E_D(\lambda, T, m = \lfloor kT \rfloor)$ decreases as k takes larger values in the interval $((v - 1)\ell, v\ell]$, that is,

$$\begin{aligned} & E_D(\lambda, T, m = \lfloor \lfloor (v - 1)\ell \rfloor + 1 \rfloor T) \\ & > E_D(\lambda, T, m = \lfloor \lfloor (v - 1)\ell \rfloor + 2 \rfloor T) \\ & > \dots \\ & > E_D(\lambda, T, m = \lfloor \lfloor v\ell \rfloor T \rfloor). \end{aligned}$$

We will proceed to show that

$$E_D(\lambda, T, m = \lfloor \lfloor v\ell \rfloor T \rfloor) \geq E_D(\lambda, T, m = \lfloor \lfloor \ell \rfloor T \rfloor)$$

for all $v \in \mathbb{Z}^+$. This is equivalent to showing that

$$\sum_{i=1}^{\lfloor v\ell \rfloor} \frac{1}{a\lfloor v\ell \rfloor - v + i} \geq \sum_{i=1}^{\lfloor \ell \rfloor} \frac{1}{a\lfloor \ell \rfloor - 1 + i}$$

for any $\ell \geq 1$, $a, v \in \mathbb{Z}^+$. According to Lemma 2, we have

$$\sum_{i=1}^{\lfloor v\ell \rfloor} \frac{1}{a\lfloor v\ell \rfloor - v + i} \geq \sum_{i=1}^{v\lfloor \ell \rfloor + v - 1} \frac{1}{a(v\lfloor \ell \rfloor + v - 1) - v + i},$$

since we can substitute ℓ with $\lfloor \ell \rfloor + \tau$, where $\tau \in [0, 1)$, which yields

$$\lfloor v\ell \rfloor = \lfloor v\lfloor \ell \rfloor + v\tau \rfloor = v\lfloor \ell \rfloor + \lfloor v\tau \rfloor \leq v\lfloor \ell \rfloor + v - 1.$$

Defining the function

$$\begin{aligned} f(a, \ell, v) &\triangleq \sum_{i=1}^{v\ell + v - 1} \frac{1}{a(v\ell + v - 1) - v + i} \\ &= H_{((a+1)(\ell+1)-1)v-(a+1)} - H_{(a(\ell+1)-1)v-a}, \end{aligned}$$

it therefore suffices to show that

$$f(a, \ell, v) \geq f(a, \ell, v=1) \quad (4)$$

for any $a, \ell, v \in \mathbb{Z}^+$.

To obtain lower and upper bounds for $f(a, \ell, v)$, we apply the following bounds for the harmonic number H_n , $n \geq 1$ [15]:

$$\underbrace{\ln\left(n + \frac{1}{2}\right) + \gamma + \frac{1}{24(n+1)^2}}_{\triangleq H_{LB}(n)} < H_n < \underbrace{\ln\left(n + \frac{1}{2}\right) + \gamma + \frac{1}{24n^2}}_{\triangleq H_{UB}(n)},$$

where γ is the Euler-Mascheroni constant. This produces the lower bound

$$\begin{aligned} f_{LB}(a, \ell, v) &\triangleq H_{LB}(((a+1)(\ell+1)-1)v-(a+1)) \\ &\quad - H_{UB}((a(\ell+1)-1)v-a), \end{aligned}$$

and the upper bound

$$\begin{aligned} f_{UB}(a, \ell, v) &\triangleq H_{UB}(((a+1)(\ell+1)-1)v-(a+1)) \\ &\quad - H_{LB}((a(\ell+1)-1)v-a), \end{aligned}$$

for $(a(\ell+1)-1)v-a \geq 1$. The lower bound $f_{LB}(a, \ell, v)$ is an increasing function of v for any $a \geq 1$, $\ell \geq 1$, $v \geq 2$, since the partial derivative $\frac{\partial}{\partial v} f_{LB}(a, \ell, v)$, which is given by

$$\begin{aligned} & \frac{2(\ell-1)}{(2((a+1)(\ell+1)-1)v-2(a+1)+1)(2(a(\ell+1)-1)v-2a+1)} \\ & + \frac{a(\ell+1)-1}{12((a(\ell+1)-1)v-a)^3} - \frac{(a+1)(\ell+1)-1}{12(((a+1)(\ell+1)-1)v-a)^3}, \end{aligned}$$

can be shown to be positive. We therefore have

$$f(a, \ell, v) \geq f_{LB}(a, \ell, v) \geq f_{LB}(a, \ell, v=2)$$

for any $v \geq 2$, $a, \ell, v \in \mathbb{Z}^+$. We now proceed to demonstrate that $f_{LB}(a, \ell, v=2) \geq f(a, \ell, v=1)$.

For the case $\ell = 1$, consider the function

$$\begin{aligned} g(a) &\triangleq f_{LB}(a, \ell=1, v=2) - f(a, \ell=1, v=1) \\ &= \ln\left(\frac{2a+1}{2a-1}\right) - \frac{81a^4 - 71a^2 + 16}{a(9a^2 - 4)^2}. \end{aligned}$$

It suffices to show that $g(a) \geq 0$ for any $a \geq 1$, which is indeed the case since

$$\lim_{a \rightarrow \infty} g(a) = 0,$$

and the derivative

$$g'(a) = -\frac{621a^6 - 961a^4 + 436a^2 - 64}{a^2(4a^2 - 1)(9a^2 - 4)^3}$$

is negative.

For the case $\ell \geq 2$, we consider the function

$$h(a, \ell) \triangleq f_{LB}(a, \ell, v=2) - f_{UB}(a, \ell, v=1),$$

which can be shown to be nonnegative for any $a \geq 1$, $\ell \geq 2$.

It follows that

$$f_{LB}(a, \ell, v=2) \geq f_{UB}(a, \ell, v=1) \geq f(a, \ell, v=1)$$

for any $\ell \geq 2$, $a, \ell \in \mathbb{Z}^+$.

Combining these results, we obtain

$$f(a, \ell, v) \geq f_{LB}(a, \ell, v) \geq f_{LB}(a, \ell, v=2) \geq f(a, \ell, v=1)$$

for any $v \geq 2$, $a, \ell, v \in \mathbb{Z}^+$, which gives us inequality (4) as required. Consequently, we have

$$E_D(\lambda, T, m = \lfloor kT \rfloor) \geq E_D(\lambda, T, m = \lfloor \lfloor \ell \rfloor T \rfloor)$$

for any $k \in \mathbb{Z}^+$. Since

$$E_D(\lambda, T, m=n) \begin{cases} = E_D(\lambda, T, m = \lfloor \lfloor \frac{n}{T} \rfloor T \rfloor) & \text{if } \frac{n}{T} \in \mathbb{Z}^+, \\ \geq E_D(\lambda, T, m = \lfloor (\lfloor \frac{n}{T} \rfloor + 1) T \rfloor) & \text{otherwise,} \end{cases}$$

we also have

$$E_D(\lambda, T, m=n) \geq E_D(\lambda, T, m=\lfloor \ell \rfloor T).$$

Therefore, if $\lfloor \ell \rfloor \leq \lfloor \frac{n}{T} \rfloor$, then $\bar{x}(n, T, m=\lfloor \ell \rfloor T)$ is an optimal symmetric allocation. On the other hand, if $\lfloor \ell \rfloor > \lfloor \frac{n}{T} \rfloor$, then we can eliminate all but the two largest candidate values for m^* in (3), since

$$\begin{aligned} E_D(\lambda, T, m=\lfloor T \rfloor) &> E_D(\lambda, T, m=\lfloor 2T \rfloor) > \dots \\ &> E_D(\lambda, T, m=\lfloor \lfloor \frac{n}{T} \rfloor T \rfloor) \end{aligned}$$

by Lemma 2. ■

REFERENCES

- [1] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4413–4430, Oct. 2006.
- [2] C. Fragouli, J.-Y. L. Boudec, and J. Widmer, "Network coding: An instant primer," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 1, pp. 63–68, Jan. 2006.
- [3] S. Jain, K. Fall, and R. Patra, "Routing in a delay tolerant network," in *Proc. ACM SIGCOMM*, Aug. 2004.
- [4] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and Wait: An efficient routing scheme for intermittently connected mobile networks," in *Proc. ACM SIGCOMM Workshop Delay-Tolerant Netw.*, Aug. 2005.
- [5] S. Acedánski, S. Deb, M. Médard, and R. Koetter, "How good is random linear coding based distributed networked storage?" in *Proc. Workshop Netw. Coding, Theory, and Appl. (NetCod)*, Apr. 2005.
- [6] A. G. Dimakis, V. Prabhakaran, and K. Ramchandran, "Ubiquitous access to distributed data in large-scale sensor networks through decentralized erasure codes," in *Proc. Int. Symp. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2005.
- [7] A. Kamra, V. Misra, J. Feldman, and D. Rubenstein, "Growth codes: Maximizing sensor network data persistence," in *Proc. ACM SIGCOMM*, Sep. 2006.
- [8] Y. Lin, B. Liang, and B. Li, "Data persistence in large-scale sensor networks with decentralized fountain codes," in *Proc. INFOCOM*, May 2007.
- [9] S. A. Aly, Z. Kong, and E. Soljanin, "Fountain codes based distributed storage algorithms for large-scale wireless sensor networks," in *Proc. ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2008.
- [10] R. Kleinberg, R. Karp, C. Papadimitriou, and E. Friedman, *Personal correspondence between R. Kleinberg and A. G. Dimakis*, Oct. 2006.
- [11] D. Leong, A. G. Dimakis, and T. Ho, "Symmetric allocations for distributed storage," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2010.
- [12] S. Jain, M. Demmer, R. Patra, and K. Fall, "Using redundancy to cope with failures in a delay tolerant network," in *Proc. ACM SIGCOMM*, Aug. 2005.
- [13] Y. Wang, S. Jain, M. Martonosi, and K. Fall, "Erasure-coding based routing for opportunistic networks," in *Proc. ACM SIGCOMM Workshop Delay-Tolerant Netw.*, Aug. 2005.
- [14] M. Piórkowski, N. Sarafijanovoc-Djukic, and M. Grossglauser, "A parsimonious model of mobile partitioned networks with clustering," in *Proc. Intl. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2009.
- [15] J. Havil, *Gamma: Exploring Euler's Constant*. Princeton, NJ: Princeton University Press, 2003.